An Architecture for Emotional Facial Expressions as Social Signals

Ruth Aylett, Christopher Ritter, Mei Yii Lim, Frank Broz, Peter E McKenna, Ingo Keller, and Gnanathusharan Rajendran

Abstract—We focus on affective architecture issues relating to the generation of expressive facial behaviour, critique approaches that treat expressive behaviour as only a mirror of internal state rather than as also a social signal and discuss the advantages of combining the two approaches. Using the FAtiMA architecture, we analyse the requirements for generating expressive behavior as social signals at both reactive and cognitive levels. We discuss how facial expressions can be generated in a dynamic fashion. We propose generic architectural mechanisms to meet these requirements based on an explicit mind-body loop and Theory of Mind (ToM) processing. A illustrative scenario is given.

Index Terms—Intelligent agents, affective computing, interactive systems, software architecture, cognitive informatics

1 Introduction

This paper poses the problem of how to incorporate a generative account of expressive behaviour into an affective architecture, focusing on facial expressions. Expressive behaviour using the body posture, gesture, glance, facial expression is an significant component of communicative content alongside the verbal channel, and is therefore required for social agents, whether robots or graphical characters. Facial expressions are considered particularly important for agents that have a face (some robots do not), since this is often the focus of glance by an interaction partner. With more than forty muscle groups [1], the face has a wide range of movements and thus substantial expressivity. It has been argued that more than half of expressive behaviour relates to facial expressions [2].

In this paper we focus on facial expressions that relate to affect. Many computational accounts, when not using scripting, treat them as a mirror of the internal affective state of the agent and as a way of signalling that state to interaction partners [3]. This makes for an architectural mechanism that is conceptually straightforward: directly connecting the affective outputs of the architecture to the expressive modalities of the agent. However it is clear that even children aged 3-4 [4], never mind adults, routinely modify their facial expressions in a number of ways related to their social context.

There is an argument for emotional transparency in a social agent. Cognitive appraisal theory suggests that emotions are generated when an event is appraised against a person's goals, with positive affect when events favour goals and negative affect when they do not. A transparent display may make a social agent's goals, and how far these are successfully met, more obvious to its interaction partner. Thus the early system Kismet [5] played the role of an infant in learning scenarios, and used transparent expressive behaviour to regulate the interaction, to encourage more or reduce the amount of stimulus it was receiving.

However many applications of embodied social agents require more sophisticated expressive behaviour. The Laura agent of [6] deliberately generated warm facial expressions so as to build trust and rapport. Where an embodied social agent aims to improve user motivation, or it operates in a training or education setting [7] then expressive behaviour is more likely to be an action explicitly chosen by the agent than a reflection of its internal state. Issues relating to long-lived interaction and getting past the novelty effect [8] also require a less naive account of expressive behaviour. Dealing with these issues is a motivation for this work.

A simple case is where a facial expression related to internal affect is muffled or suppressed. A classic example would be playing the card game poker, though a subordinate being reprimanded by their boss, or a parent walking with their child on a dark night, might be expected to suppress expressions of anger and fear respectively. In the case of poker, there are game-related social norms; in the other two examples a mix of social norm and in-situ estimation of the impact of ones expressive behaviour on others in the shared context.

[9] distinguishes four categories of expression modification (*display rules*): the cultural; the personal (depending on

- *R. Aylett is with MACS, Heriot-Watt University, Edinburgh EH14 4AS, United Kingdom. E-mail: r.s.aylett@hw.ac.uk.*
- *C. Ritter is with CITEC, Bieldefeld University, Bieldefeld 33615, Germany. E-mail: critter@techfak.uni-bielefeld.de.*
- *M.Y. Lim, F. Broz, and I. Keller are with the School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh EH14 4AS, United Kingdom. E-mail: {M.Lim, f.broz, ijk1}@hw.ac.uk.*
- *P.E. McKenna and G. Rajendran are with the Department of Psychology, Heriot-Watt University, Edinburgh EH14 4AS, United Kingdom. E-mail: {p.mckenna, T.Rajendran}@hw.ac.uk.*

Manuscript received 18 June 2018, revised 6 Mar. 2019, accepted 12 Mar. 2019, Date of publication 0 . 0000; date of current version 0 . 0000. (Corresponding author: Ruth Aylett.)
Recommended for acceptance by C. M. de Melo, C. Becker-Asano, D. C. Moffat, J. Parthemor, and D. D. Petters.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2019.2906200

personality or other individual factors); vocational requirements (as in actors); and the need of the moment. Giving a social agent this ability to suppress expressions requires a process tracking the contextual impact of a given expression. Empathy, in which the observer responds to the affective state of another is one means of carrying out this tracking. More generally, one might posit Theory of Mind (ToM) capabilities, in which one takes into account how one is likely to be perceived by another [10]. At the very least the agent must be able to recognise and adhere to social norms for a given situation.

In the pedagogical example above, expressive behaviour may also be generated for a specific communicative objective. Likewise, economic games used to study negotiation may deploy expressive behaviour competitively as game moves [11] or as an aid to reaching cooperative decisions [12]. The classic case of an unwelcome birthday present may result in a deliberate facial expression of pleasure in order to convey gratitude. As cited above, children as young as 3 or 4 perform this modification. Indeed smiles are notoriously ambiguous about internal state and very often related to social context [13]. A recent account [14] gives three types of social smiles: those rewarding the behaviour of others, as in this example, those creating or strengthening affiliative social bonds, and those regulating social hierarchies. We should note that embodied social agents without specific expressive behaviour may still have their behaviour treated as socially expressive. [3] examples a robot that turned away from a user immediately following a request and was interpreted as showing dislike or contempt. Thus an embodied social agent that can predict the social impact of its expressive behaviour may help to prevent misunderstandings. We propose to do this through an explicit mind-body loop and the application of ToM capabilities.

The key point is that facial expressions operate as social signals not merely as information about internal state. Even the greater emotional transparency of infants relates to a social context where carers are motivated by smiles and will act to deal with the causes of negative affective states. We argue that coupling the affective outputs of an agents architecture to its expressive modalities is insufficient in the development of embodied social agents.

By modelling the ability to handle expressive behaviour as a social signal, we broaden the range of applications to which a social agent can be applied and offer a standard mechanism for regulating expressive behaviour - whether by suppression or substitution - rather than ad hoc application-dependent solutions. We also broaden its communicative repertoire by explicitly including expressive behaviour in the set of actions from which it can select rather than binding it to the internal affective state being modelled. We seek to retain some of the interactional advantages of transparent expressive behaviour by supporting modification and not just overlaying of expressions, allowing micro-expressions [15] that we refer to elsewhere as the *Partial Poker Face* [16].

However, should one equip social agents with what could amount to deceptive behaviour? There are ethical considerations in creating convincing liars even though we know that many humans behave like that [17]. But the line between actually lying and what we describe as social facilitation is very blurred. [18] argue that up to 30 percent of social interactions of longer than 10 minutes contain deceptions about affective state. Thus more long-lived social agents do need more sophisticated expressive behaviour to create smooth interaction.

A second set of architectural requirements is raised by the dynamic nature of expressive behaviour. Some representations, such as the Facial Action Coding System (FACS) [1], define specific static facial expressions. In reality, facial expressions are nearly always continuous and dynamically varying, along with the speech stream they often accompany. State-based expressive behaviour fits well with a state-based architecture, with dynamics confined to interpolating between the defined expressive states. In addition, a state-based approach fits well with explicit representations of affective state, where a dynamic approach to expressions is more consistent with a process-based architecture and implicit representations. We return to these issues in the next section.

2 Background

2.1 Theoretical Issues

While we have argued above against wholly transparent affective expressive behaviour, the idea that expressive behaviour represents affective state in humans at all is far from uncontested. [19] argues strongly the *behavioural ecological view*, that facial expressions are not related to affective state but are entirely social signals produced by an evolutionary process. However we do not align with this more radical viewpoint, being more convinced by arguments against it in [20] and studies such as [21] which shows widespread interpretation of facial expressions as indicative of affective state.

A more categorical position still is to reject the idea that affective states cause actions at all, not just expressive behaviour conceived as action. This relates to discussions of Basic Emotion Theory (BET), that a finite set of emotions such as anger, fear, disgust, happiness, sadness, surprise and possibly others, emerge from evolutionary processes related to survival and operate reflex behaviours. As [22] argues in relation to his New BET, discussion is bedeviled by using linguistic labels to mean different things, from processes at different levels of abstraction (e.g., physiological sensations v cognitive categories) to different affective states (are all forms of *anger* the same?). In this paper we take the perspective of cognitive appraisal theory, that affective state creates action tendencies priming actions rather than inevitably producing them, while we also model lower-level processes that have the character of reactions, if not reflexes.

A generic issue is how far an affective architecture can be considered social rather than merely individual. Computational architectures based on psychological theory tend to import individualist assumptions. The *Big Five* personality dimensions [23] sometimes used for behaviour generation in embodied social agents focuses on individual patterns of behaviour. Cognitive appraisal theory [24] is not per se incompatible with the modelling of social and cultural processes, but its focus on interaction between external events (stimuli) and individual goals prioritises the individual. Where the goals are taken as a given, affect will then represent an established individual reaction to the social context.

Indeed, [24] does not distinguish between appraisals relating to an event and to a person, so that socially-determined emotions such as *sorry-for* (someone) are modelled in exactly

the same way as the *fear* generated by a threat to one's survival. Yet the social context is known to have a substantial effect on individual appraisal: [25] gives the example of watching a comedy you enjoy with a close friend who disapproves of it.

Social appraisal [25], [26] involves appraising the thoughts or feelings of others, especially those with whom there is a relationship, as well as the emotion-causing event itself. This view stresses the importance of empathic reactions and the role of expressive behaviour in social regulation processes which might result in modifications of expressive behaviour. Note that social appraisal does not require an actual change in internal affective state. Sensing the disapproval of a friend - a social signal - one might actually find a comedy less amusing, or, for affiliative reasons, suppress the expression of one's amusement.

Social appraisal is not unlike the idea of coping behaviour [27]. Coping behaviour, a reaction to an affective state, has an external path in which actions-in-the-world are carried out to make the world more compatible with the goals of the individual. It also has an internal path, where cognitive strategies adjust a painful affective state, for example by perceiving a 'silver lining' to an unpleasant event. Internal coping behaviours are a second possible source of expressive behaviour modification.

Cognitive appraisal architectures can be extended into a more socially responsive form by adding explicit mechanisms to handle social interaction. Thus the FAtiMA architecture [28] has been extended with a simulation ToM capability [29] (see Section 4), and the ability to model culturally-specific behaviour [30]. Both offer mechanisms supporting the modification of expressive behaviour.

The developmental robotics approach [31] is more likely to give the social context priority since it considers the construction of internal architectural structures by interactional processes such as *enaction* [32]. However since this work is driven by the analysis of very young infants, most of that looking at communication considers basic capabilities like mutual glance and the development of turn-taking. For facial expressions [33] the issue of interest is how to learn a mapping between expressions and internal state. This does not bear on the problem considered here.

A further theoretical dimension is a static versus a dynamic account of behaviour, with implications for the approach a computational implementation must take. In its early form, cognitive appraisal was distinctly state-based: an event was compared with the goals of the individual, a set (usually) of labelled emotions was generated to enable action tendencies. More modern versions of cognitive appraisal, such as the Component Process Model (CPM) [34], break appraisal into a sequence of related actions in a pipeline of evaluation phases, each with a set of subchecks that may be shared between phases. The main phases of the CPM concern *Relevance*, *Implication*, *Coping Potential*, *Normal Significance* where the last of these refers to social norms. Linking these to Facial Action Units, as CPM does in some cases, allows facial expressions to be generated directly from the appraisal process [35] without having to pass through labelled emotions, thus producing a more dynamic model.

A multi-stage model reduces the granularity of each step, moving in the direction of process. In addition, because facial expressions can be driven at different stages on different timescales, it supports micro-expressions and expression modification. However multi-stage appraisal is not the only way of dealing with this issue. A different class of models, those based on drives and homeostasis, are more directly process-based [36]. The PSI model [37] works with drives based on five basic needs: personal survival (food etc); species survival (sex etc); affiliation (belonging to a group, engaging in social interactions); certainty (the need for predictability of events and consequences) and competence (the need to master problems and tasks, including meeting needs). Drive-based architectures work with upper and lower bounds on needs, setting a *comfort zone* within which the values are acceptable. If a need moves out of the comfort zone, the drive seeks to activate behaviours to move it back - this is the process of homeostasis which is inherently dynamic.

The PSI model has no direct representation of emotion, but the behaviours generated by the drives are interpretable as having affective qualities such as anger, joy or anxiety. It outputs numerical values of valence and arousal which can be used to synthesise multiple expressive behaviours without having to pass through labelled emotions or necessarily through facial action units [38]. The downside of a model at this lower level of abstraction is that it is difficult to use for embodied social agents using natural language, since language by definition works at the symbolic level. This suggests a multi-level model, very common in robotics, in which moment-to-moment behaviour is controlled by drives but strategic decisions are made via appraisal and planning. This is the approach taken in the FAtiMA architecture discussed below. A multi-level model is also a multi-stage model, with lower layers typically working on shorter timescales, thus also supporting micro-expressions and expression modification.

2.2 Implementations

In this section we consider implemented systems that deal with expression modification and also with facial expressions generated as explicit communications.

One body of work on expression modification is concerned with combining more than one emotion to produce mixed facial expressions, for example an immediately generated emotion with longer term affective states like mood [39]. This is however still a version of transparency. [40] discusses the issue of social modification of expressions but focuses on the actual composition process. [41] takes this idea further by considering how different emotions might arise from an egocentric appraisal and an empathic appraisal and evaluating the impact on users of *masking* (empathic expression overrides egocentric expression) and *superimposition* (expressions are combined). However expressions were hand-coded rather than generated autonomously by an architecture and the focus was on realising and then evaluating the expressions in a graphical character.

Empathic behaviour requires facial modification within a social context. It is a significant issue in work on pedagogical agents, though some work [42], [43] relates to natural language expressions rather than non-verbal behaviour. [7] discusses robot expressive behaviour in a tutoring context, but the robot used had no facial expressivity and relied on gesture, while affective responses were derived from an

application-specific learner model that would not generalise to other domains.

[44], in a role-play therapeutic domain, argues expressive behaviour may relate to affect generated by an agent's own coping actions, citing guilt as a result of a shift-blame action. This produces a sequence of expressive behaviours but still relates to the actual affective states of the agent. This application couples a cognitive architecture similar to the one we modify to a pre-authored dialogue model; as it involves agent-agent rather than human-agent dialogue, it can be certain about the affective states each agent is responding to. Thus it combines a cognitive architecture able to model a rich internal state with the focus on interactivity of a dialogue system, albeit a pre-authored one.

Interesting work on affect in interaction has been carried out in the context of negotiation games. [11], [45] studies the social impact of an agent's display of joy, sadness, anger and guilt, and how they function as social signals. These studies support learning of the parameters for a Bayesian network giving probabilistic predictions, from emotion displays, of how the negotiating partner appraises the interaction. This supports predictions about their intentions [46]. This is not intended to be and is not a generic architecture but is specific to the iterated prisoner's dilemma used in the studies.

Other work on expressive behaviour as social signal focuses on deception and lies. [17] investigates how deceptive expressive behaviour may be used to produce a desired outcome in an economic game, using a similar approach to [45]. Focusing on a particular element of negotiation, this work demonstrates that agents with a deceptive facial expression when they make an offer do attain the desired negotiation outcome. However the study was carried out with video recordings of agent expressions rather than with a generating architecture.

[47] directly investigates expressive behaviour for an agent telling lies, building on [48]. It argues that facial expressions will not be completely deceptive because some facial muscles are not controllable at the conscious level. Thus both micro-expressions and compound facial expressions will occur. This work also considers timing, with faked expressions lasting longer than transparent ones and asymmetry, where faked expressions have more activity on one side of the face than the other. Two studies were carried out, which both used smiles to mask other expressions in a similar way to [40]. In the first study, smiles were either straightforwardly happy or combined with disgust: these conditions were hand-coded.

A second study used a liar dice game in which lies are part of game play. As with other work using games, the context is easy to assess, depending entirely on the game play, so a generic architecture for deception was not required. In both cases, the aim was to evaluate the social impact of the compound expressions.

Work that is very relevant to this paper came from [49]. This distinguished between an impulsive agent one that directly expresses its affective state and a reflexive agent, which could refrain from expressing its state. However it was based on the annotation of dialogue plans rather than an affective architecture. These were held in a plan library within the framework of a BDI architecture [50] originally designed for rational agents and only later extended, in conceptual form, to incorporate affect [51].

This view of expressive behaviour as a multi-modal adjunct to language communication comes from a community with a different focus from cognitive or affective modelling. It generalises the idea of a performative language action into non-verbal behaviour [52]. It has a strong focus on interactivity, but the social signal aspects of expressive behaviour are inferred from the dialogue moves with which it is associated. In this tradition, dialogue is viewed as a means of changing beliefs in a purely logical model [53], an orthogonal view to cognitive modelling. It supports affective communication but without a modelled affective state or any affect-generating process.

Cognitive model-based research puts language actions on a similar level to other agent actions rather than giving them control of agent activity, while in dialogue system research, agent actions are determined by a dialogue manager. This delegates the control of expressive behaviour to a process that annotates utterances using a mark-up formalism ([54], [55]).

Mark-up of a dialogue stream both gives primacy to language over expressive behaviour, and assumes that the decisions about what to communicate are made by the Dialogue Manager before affective expressive behaviour is generated. However a social signal approach requires that affective choices be made at the level of action selection in the architecture. We also argue that the modification process requires an internal circuit within the architecture, since if the agent does not know what affective response it has chosen, it cannot modify it in a contextually sensitive manner. In human terms, you need to know you are angry in order to suppress anger.

This follows the work of [56] and in turn the ideas of [57] who stresses the somatic aspects of emotion, an emotional body state, which feeds into later processing at a more cognitive level. It is also consistent with the more sophisticated view of cognitive appraisal already mentioned [58] as a multi-stage process with a temporal profile, consisting of cognitive appraisal, a physiological activation and involving arousal, motor expression (expressive behaviour), a motivational component, and a state of subjective feeling. If one sees expressive behaviour as integral to emotion in this way, then even as a reflection of internal state it poses architectural issues.

Finally, [59] addresses some of the same questions as this work but focuses on social signal analysis so as to establish the affective state of an interaction partner rather than social signal generation. It is concerned with recognising the social signals associated with emotional regulation - or coping behaviour - focusing on *shame*, generated in a cognitive appraisal framework as a result of agent actions that have negative praiseworthiness. It incorporates a simulation model of ToM, similar to that discussed below in Section 4, implementing a shame model in its own architecture to make predictions about an interaction partner's likely social signals, and thus aid Bayesian recognisers in detecting them. This maintains a transparency approach. It does enrich its model by including the target of the expressive behaviour - an issue not yet dealt with in the work here.

3 Representational Issues

3.1 Architectures

All computational architectures are shaped by their representational choices. We have already referred to one significant dimension, the choice between state-based and

```
vrAgentBML request Alice Bob SPEECH_24
<?xml version="1.0" encoding="UTF-8"?>
<act>  <bml>
  <speech id="b1">
    <text>
    Wait! That's <tm id="not"> not what I mean.
    </text>
  </speech>
  <sbm:interrupt id="b2" act="SPEECH_23"
                 start="b1:start"  />
  <gaze id="b3" target="Bob" start="b2:start" />
  <head id="b4" type="SHAKE" stroke="b1:not" />
</bml> </act>
```

Fig. 1. An example of BML from SmartBody, controlling Glance.

process-based representations. In affective architectures, this has tended to result in a choice between symbolic representations actuated through inferencing (for example [28], [39] and non-symbolic representations in which homeostasis is a dominant mechanism (for example [37]).

Dialogue management systems were once symbolic systems, manipulating natural language. More recently, after the success of statistical approaches in speech recognition, machine learning on large corpuses of spoken dialogue in specific domains has encoded probabilistic transitions between dialogue actions [60]. This makes the addition of expressive behaviour, especially that related to affect, more difficult as it requires analysis of the chosen dialogue action with fallible approaches such as sentiment analysis. It produces an architecture in which the role of inferencing is substituted by transitions in the learned network, an implicit encoding. Thus the issue of explicit versus implicit representation is a further dimension.

The most important representational decision in an affective architecture is how to represent affect itself, determined in large part by the chosen theoretical framework. A simple state-based architecture may represent affect as a single symbolic variable and an associated intensity value, as in the OCC model [24] with its 42 named emotions, while a model built around drives and homeostasis may have no explicit representation of affect at all [37] but generate affective behaviour as dynamic patterns. A multi-stage theory such as the one in this work may involve multiple representations of affect, in particular if affect is seen as a phenomenon on more than one architectural level.

The FAtiMA architecture used as the basis for the ideas below [28] divides into a reactive and a predictive component working on different time-scales and controlling different types of behaviour. Its predictive component runs a planning system, which is where overt communicative intent would be handled. But some behaviour cannot reasonably be thought of as consciously planned - take the example of bursting into tears at the death of a loved one. FAtiMA incorporates a reactive layer that triggers much shorter-term unplanned behaviours. Note that incorporating a reactive system does not in itself force the choice between a symbolic or non-symbolic representation since symbolic rules can play this role.

We have seen that cognitive appraisal itself may be decomposed into multiple stages suggesting a collection of processes rather than a single process. There is physiological evidence [61] of different brain mechanisms being involved in what is known as emotional or affective empathy (or sometimes as emotional contagion), and cognitive empathy, based on reasoning about the affective state of another.

Damasio [57] distinguishes between primary and secondary emotions. The former are seen as innate, relating to fast and reactive behaviour patterns such as fight/flight, or infant distress behaviours, that do not involve cognitive-level processing, and are closely tied to specific stimuli. Secondary emotions like *admiration* or *hope* are ascribed to higher cognitive processes involving expectations, learned outcomes and social context. Note that this distinction does not correspond exactly to the language labels that we may use: *fear* may count as a primary emotion if one is attacked by a ferocious dog, but we may use the same label in relation to an event that has not yet happened, an inverse to hope.

Primary emotions independent of the social context, are good examples of emotions an agent might suppress after a later evaluation. If a ferocious dog attacked while one was taking a child for a walk, a flight response activated by a primary emotion of fear might be suppressed in favour of an attack on the dog to defend the child. The language label for this would be *courage*. However, the anger one might feel witnessing the action of a bullying boss against a fellow employee is not primary by this definition, though its expression or suppression is also subject to evaluation of the social context.

While the distinction between primary and secondary emotions is not wholly useful for this work, that between somatic and cognitive impact does capture the stages to be modelled so that an agent can *feel* an emotion so as to modify it. The modelling issue is how to represent affect in cognitive and somatic systems and how to link these representations both to evaluation of social context and to the generation of expressive behaviour.

3.2 Representations for Facial Expressions

Accounts of expressive behaviour that reduce it to multimodal annotation of the output from an affective architecture (or indeed, from a dialogue manager [62]) pose the representational problem as one of mapping from the architecture (conceived as mind) to the agents actuation capabilities (conceived as body). Interesting work in graphical characters has moved towards a standardised mark-up language for this purpose, Behavioural Markup Language (BML) [63] and to middleware based on this [64] such as SmartBody. Fig. 1 gives an example of BML controlling an agents gaze in SmartBody [64].

The standard BML flow assumes that behaviour planning will deal with annotations on utterances from a higher-level process. It leaves no space for a somatic representation that can reflect emotion back into the cognitive system to be re-evaluated. Such a causal chain would run as shown in Fig. 2 producing a mind-body-mind loop.

The somatic level both dispatches output to the actual generation of expressive behaviour and is also evaluated so that the agent *knows what it is feeling* and is able to start the modification process. The somatic level also supports modelling of involuntary expressive changes - for example blushing or crying. The temporal overlap produced by

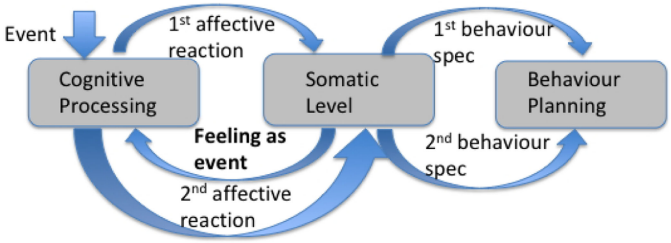

Fig. 2. Affect re-evaluation.

re-evaluation is consistent with the idea of micro-expressions [15] that facial expressions will reflect internal affective state for only a very short time before being replaced by the socially determined expression. This idea is central to our proposed architecture.

A somatic level also has a role to play in mapping affect onto an agent's expressive capabilities. Happiness can be expressed as a smile if a social agent can smile. If not - for example a robot with no face, or a face without a moveable mouth - other modalities can be selected.

Work has taken place to refine the markup system and to add a specific virtual body representation [65] in the Thalamus system. This inserted a BodyInterface unit between the agent mind and behaviour planning in the same way as the somatic level in Fig. 2. It has the ability to both receive and send messages. This incorporates a feedback mechanism, needed for the expressive capabilities under discussion, but, as conceived, deals with external events and not internal affective events.

The two most widely-used systems for transforming an affective response into a behaviour specification in embodied agents are the Facial Action Coding System [1] and dimensional systems, of which the Pleasure-Arousal-Dominance System (PADS) [66] is the most popular.

FACS defines 44 muscle groups on the human face and relates muscles to expressions via facial Action Units (AUs), specific configurations of these muscle groups. It is often used by researchers who want to generate facial expressions in social agents, but is of much wider applicability. It was designed for facial analysis, for example on videos, and is still much more widely used for this purpose than for generation. It is also used for research into expression recognition.

Used generatively, FACS offers a way of defining certain facial expressions with respect to specific affective states, and a tool is available for doing this [67]. In particular, AUs can be used to define Ekman's (contested) conception of primitive emotions [68] facial expressions corresponding to fear, anger, disgust, happiness, sadness, surprise that are said to be recognised across cultures (though this has been recently challenged; see [69]). However, if these primitive emotions are seen as comparable to the primary emotions discussed above, then they are targets for modification, and indeed they are rarely visible in everyday adult social interaction. They thus form a basis for blending, as in [47] discussed above.

An AU-defined facial expression representing an affective state is straightforward to interface to an affective architecture outputting such states. However, this is also a disadvantage, since it produces a static and rigid mapping. The very concept of an *expression*, as against a behaviour, works poorly in actual interaction as distinct from in a photograph or a video frame. AU definitions say nothing about how the face moves into and out of an expression.

FACS can be used in a more dynamic manner. The decomposed appraisal process of [58] discussed above associates AU changes with its various stages. This has been implemented in some embodied social agents: [35] applied Sherer's theoretical framework in a game environment. It used only the labelled emotions Joy, Sadness, Guilt, Anger alongside intermediate appraisal step changes, but there is no evaluation detailed. [70] used *Hot Anger* and *Fear*, but found many questions about dynamics unanswered by the theory. [13] investigated user perceptions of different smiles. It found these were impacted by issues such as amplitude, duration, onset and offset velocities and not just the AU. Recent work in robot expression recognition by the authors also concluded that the dynamics were very important in recognition [71].

A final point is that even humanoid social agents usually lack equivalents the full set of Action Units, with faces much less expressive than a human's; even more true for robots. Subtle affective behaviours van still be produced if alternatives are sought - which may draw on film animation as well as psychological theory.

Dimensional systems offer a numerical representation of emotion in a space defined by two or more dimensional axes but do not directly provide expressive behaviour. Indeed, emotions can be represented as locations in a numerical Pleasure-Arousal-Dominance space (the PAD system), symbolic labels attached to locations, and then used to drive AUs. A more consistent approach to facial expressions in architectures using PAD would involve driving facial features (or AUs) directly from the dimensional values. Here it is not necessary to translate to a symbolic affective label and then back to numerically-driven motor action.

An example of work taking this approach, though using Pleasure (Valence) and Arousal only, and not Dominance, is that by Lim and Aylett [38]. This applied the drive-based PSI architecture [37] in the context of a story-telling guide, and directly linked the output valence and arousal values to a simple 2D graphical face with limited expressive features. In the absence of a single psychological theory linking valence and arousal to specific facial features, this work adopted a number of heuristics found in the literature affecting eyes, brows and mouth, thus bypassing linguistic labels for emotion. Fig. 3 shows part of the resulting facial feature space for valence against high arousal (in this system high arousal was 0 and low was 1).

The advantage of driving expressive behaviour like this within a process based architecture, is that behaviour will be naturally dynamic, and the issues of merging different expressive modalities are dealt with separately for each feature. The disadvantage is that it lacks the experimentally-validated status of FACS.

If the somatic representation in use is not a symbolic one, using PAD space to transform numerical triplets (P, A, D) into a symbolic representation of affect allows the somatic representation to be manipulated much more easily in cognitive-level processing. This is useful since we will see that social interaction theories are easier to represent symbolically. A reactive system using drives can produce a rapid affective output, using PAD space, feeding directly into motor action, at the same time as outputting the nearest

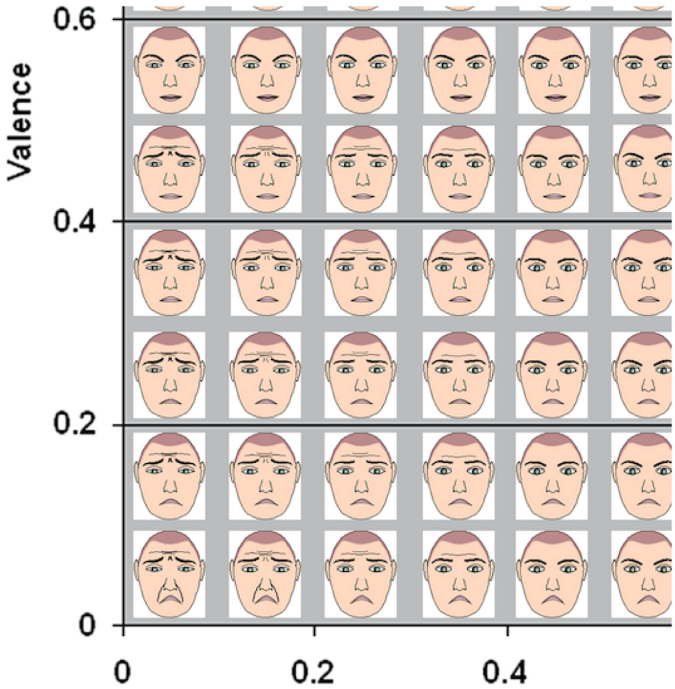

Fig. 3. Facial expressions driven directly by valence and arousal.

symbolic label in the space to cognitive processing. Though this is not how our illustrative architecture below works, we point out in our conclusions that there may be very good reasons for taking this approach.

4 Base Architecture

Including a mind-body-mind loop for expressive behaviour is independent of the detail of the architecture used. Any architecture that represents mind and body components and a structural inter-connection between them could take up these ideas with different implementational details.

However the approach requires an architecture modelling social interaction so that feelings returned to for re-evaluation can be assessed in the social context. It also requires a ToM mechanism able to assess expressed feelings for their impact on the interaction partner. There are few existing architectures to choose from. Building a new architecture from scratch is certainly possible, but in this paper the aim is to explain clearly how to deal with expressive facial behaviour as a social signal, so using an existing architecture reduces the size of the task.

For this reason, we start from the FatiMA architecture already mentioned [28] a cognitive appraisal architecture which has already been extended with social interaction functionalities [72], in particular the Social Importance Model of Kemper [73] - see Fig. 4 and a simulation-based Theory of Mind capability [29].

FAtiMA deals with events along two time-scales: a reactive timescale without any intermediate processing, and a deliberative timescale that allows for planning or other cognitive processing before selecting an action. These can be seen in Fig. 4 to the right top and right bottom. The model also includes a Memory component in which KB is a knowledge-base of the surrounding world and its objects, and AM is an affective memory of past interactions supporting mood modifications. The motivational state contains goals, and activated goals/current intentions.

The reactive system is required for immediate expressive behaviour unrelated to planning expressions of intense distress such as crying, or of involuntary laughing. Architectures that only implement affective transparency could deal with the whole of expressive behaviour like this. However, an advantage of using FAtiMA to discuss social signals is that it also generates affective responses via its Deliberative Layer supporting planned or other cognitively processed expressive behaviour as well.

Neither layer in the existing architecture entirely captures the issue under discussion of modifying expressive behaviour. We argue that modification can both act as a Deliberative Layer activity and as a reaction to the agents internal feeling of its affective state.

4.1 Social Importance Model

The Social Importance Model [72] is an example of integrating social rules into a cognitive appraisal model. Kemper [73] focuses specifically on power and status, both of which are contribute to the modification of expressive behaviour for three of the categories categories cited by Ekman [9] culture,

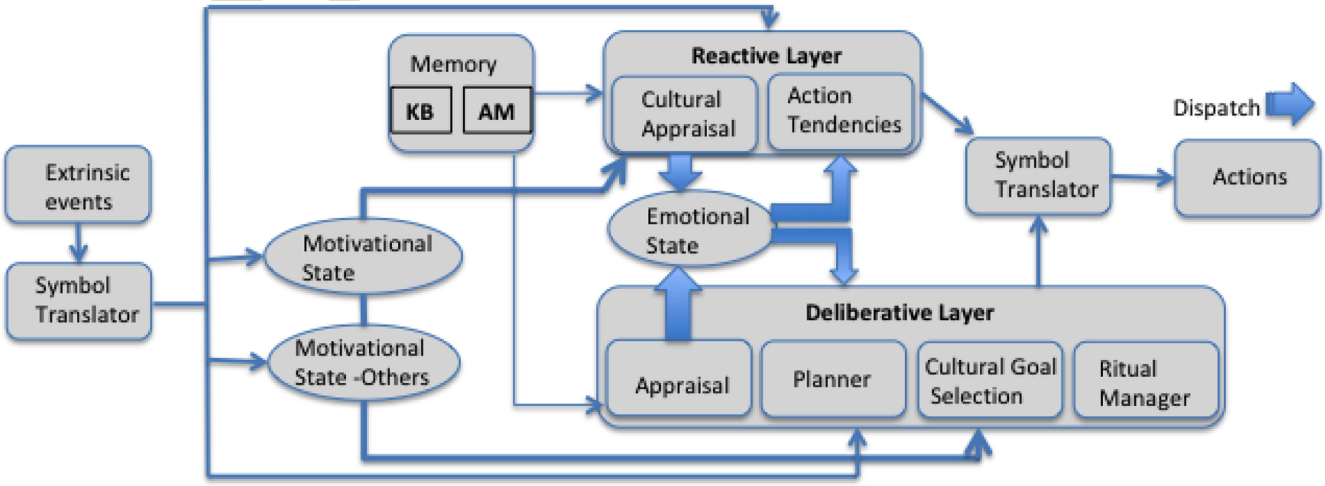

Fig. 4. FAtiMA with Kemper modelling.

vocational, and needs of the moment. We here summarise the implementation and refer the reader to [72] for further detail.

In the implementated system of 4, Social Importance (SI) represents an aggregated generalisation on the intuitive meaning of status, since the SI an agent is willing to ascribe to another may be influenced by inter-personal liking, group membership, adherence to social norms, expertise, and personal attributes such as wealth or strength. SI is not seen as a static quantity but can be increased or decreased during the course of social interactions.

The model contains three types of rules in its Reactive Cultural Appraisal function: SI Attributions, SI Conferrals and SI Claims. An Attribution occurs when an agent meets another agent and uses social rules to decide how much SI it should have. Conferrals are associated with agent goals and result in actions that acknowledge through behaviour the SI attribution an agent has given another. Expressive behaviour is one example of a conferral mechanism: as in the example of looking pleased at a birthday gift even if the agent does not like it. Conversely, an SI Claim is behaviour carried out by the agent to assert its own SI in the eyes of another agent, determinable using the Theory of Mind system discussed in the next section.

This architecture also includes a component for dealing with items that are socially symbolic rather than merely functional. Examples include wearing specific clothing like evening dress, or presenting a bouquet of flowers to a soloist at the end of a concert. Such items impact both on the agents motivations and its model of the motivations of others. This creates extra inputs into Goal Selection in the Deliberative layer. Finally the architecture can store specific plans relating to social rituals. These are defined as action sequences with a specific social meaning that must be executed with a fixed order and content - for example greeting someone, whether by shaking hands, bowing or kissing cheeks.

Relative SI has an obvious role in the modification of expressive behaviour. If another agent has a very high level of SI, then an agent is likely to suppress negative expressive behaviour such as anger or distress. If two agents have equivalent levels of SI as in two close friends then much less modification of expressive behaviour is required.

Note that this SI model has its own input into expressive behaviour as with other aspects of an agents internal processes. Social signals of disapproval or embarrassment could be invoked by another claiming more SI than an agent has attributed to them, while approval could be invoked by another attributing the SI an agent has attributed to itself. In these cases the agent generates a negative or positive affective state but the extent to which this is expressed will depend on the relative SIs involved

4.2 Theory of Mind

[10] defined Theory of Mind as the ability to infer the full range of epistemic mental states of others, i.e., beliefs, desires, intentions and knowledge. The abstractions we make about the states of mind of others and consequently of our own, is a mechanism that helps to make sense of their behaviour in specific contexts and predict their next action. A single-level theory of mind allows us to represent an embodied agents beliefs about another embodied agents beliefs and is the minimum needed to consider the impact

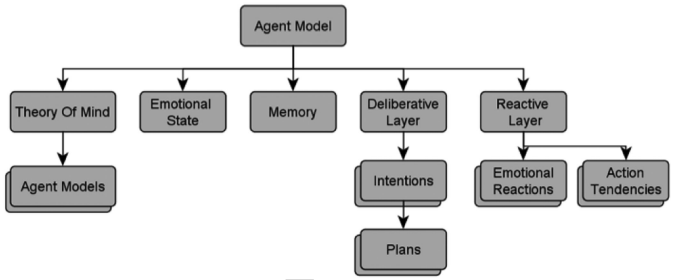

Fig. 5. Simulation of ToM through recursion.

of one's own expressive behaviour on someone else. Most adults have at least a two-level ToM allowing them to think about beliefs about another's beliefs about another's belief's - who might well be you.

There are two conceptually different approaches to the human theory of mind: the Theory-Theory approach (TT) and the Simulation-Theory approach (ST). According to TT, the mental state we attribute to others is not observable, but is knowable through intuition and insight. In implementation, this is achieved by using inference rules to reason about the beliefs of others over an explicit model of the other.

On the other hand, ST claims that every person simulates being another while trying to reason about their epistemic state. This means that one can use the same structures and processes used to update ones own beliefs and knowledge to simulate those of others. In implementation, this involves re-running the agent architecture as if for a different agent, and this is conceptually straightforward for a cognitive appraisal architecture such as FAtiMA [29].

Let us assume that Agent1 (A1) is the agent carrying out the ToM evaluation and Agent2 (A2) is the target of this ToM. Then in general for an action X1 of A1 :

1) set X1 to be the event E1 for appraisal
2) Run a copy of A1 on E1
3) Take X2 output by this new appraisal as the action of A2

This recursive use of the agent architecture to simulate ToM is shown in Fig. 5.

In this form A1 assumes A2 is exactly the same as themselves. However if one also applies the Social Importance model just discussed with A1 and A2 interchanged, then effectively the ToM is modified to take into account the difference in the social relationship from the point of view of A2.

This work was implemented [74] as part of a group-based deception game, Werewolves, in which one agent is secretly a were-wolf, able to kill other players in a segment where everyone has their eyes closed. After the event, agents take it in turns to accuse each other, and it is clear that the agent playing the werewolf not only has to lie about their status but accompany it with convincing expressive behaviour if it is to play well.

5 Updated Architecture

A number of updates to this architecture are needed in order to add the capability of modifiable expressive behaviour. Here initial conceptual work has been carried out under the banner of the Partial Poker Face [75] capturing

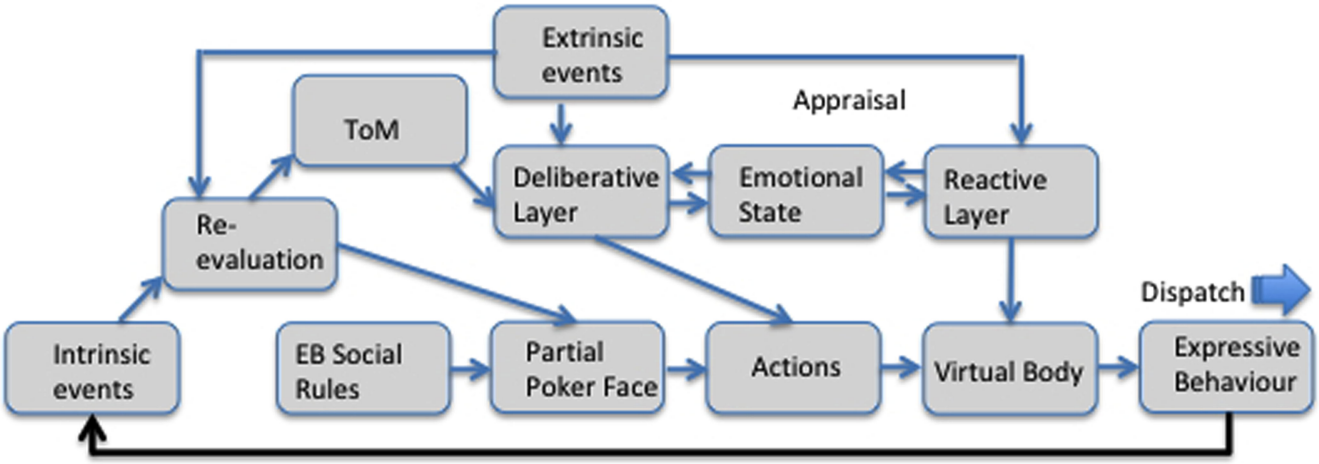

Fig. 6. FAtiMA modified to allow modified expressive behaviour.

the idea that expressive behaviour modification in humans is rarely perfect. Fig. 6 shows the changes that would have to be made to the FAtiMA architecture.

This is a slightly simplified version of 4 with some additions: Intrinsic Events, Virtual Body and Expressive Behaviour components, Partial Poker Face (PPF) and Expressive Behaviour (EB) Social Rules, Re-evaluation and ToM. Actions are planned sequences from the Deliberative Layer, Partial Poker Face and EB Social Rules are in fact part of the Reactive Layer, extracted here to make the diagram clearer.

In order to motivate these changes, we work through a scenario from [72] used to discuss the SI model of Section 4.1. In this scenario, a traveller enters a bar after failing to find the way to their hotel. At the start of the scene, there are only two characters sitting in the bar and they are talking to each other. The barman is absent (although he later appears). The goal of the traveller is to find directions to their hotel. In the version discussed in [72], the traveller is an avatar directed by a human user, and the discussion focused on the behaviour of the two agents in the bar. We adapt it to investigate how the expressive behavior of a social agent version of the traveller would be generated in the proposed architecture.

5.1 Intrinsic and Extrinsic Events

The first change that is needed to construct the mind-body-mind circuit at issue is a distinction between events external to the agent entering into cognitive appraisal (extrinsic) and events within the agent (intrinsic) generated by its affective responses. This distinction was not made in the ToM discussion above because the original motivation for that work was allowing an agent to evaluate the impact on others of its external actions upon the world. Expressive behaviour resulting from affective responses in agents exhibiting transparency had so far been considered to be hard-wired. As social signals they can be thought of as responses to intrinsic events. Note that intrinsic events may not be entirely invisible to other agents where they are associated with truly involuntary behaviour at the physiological level, such as blushing or sweating.

In the scenario, the traveller asks the bar agents if they can give directions to the traveller's hotel. In the discussion of [72], if these agents come from a collectivist culture, they will be offended by the request since the traveller is not in their in-group, and they will scowl and tell the traveller to wait for the bar man to arrive.

In our simulation, the traveller agent will then appraise the rejection of their request as a goal failure, and within the reactive layer interpret the scowls as disapproval. This will generate a negative emotion of anger which is dispatched to the Virtual Body. This corresponds to the somatic component of 2. The associated expressive behaviour is sent out to the expressive system but the feeling of anger from the Virtual Body is passed back to the Intrinsic Event (IE) component, tagged with the Extrinsic Event (EE) ID and time-stamp that originated it. This raises an intrinsic event, with an ID and time-stamp, making the agent aware of its emotion, thus beginning the mind-body-mind loop.

5.2 Re-Evaluation and Partial Poker Face

Conceptually, modifiable expressive behaviour must operate on a rapid reactive level as well as on a slower deliberative level. Otherwise modification would occur quite late and the underlying emotion would result in clearly identifiable expressive behaviour. The speed of this reactive layer is a variable in the personal presentation of an embodied agent - some agents might suppress initial facial expressions much faster than others, just as the intensity of the initial emotion might vary between individual agents too and be therefore harder or easier to suppress.

Thus the event signalling the traveller's anger is fed back through the re-evaluation module to the PPF component which is a set of reactive rules about dealing with emotions. These rules draw on the agent's knowledge of social expressive behaviour which are flattened versions of ToM reasoning - that is to say compile the output of ToM reasoning into simple rules. In this case, PPF draws on a social rule that says if an agent expresses anger to someone it causes an angry response with high undesirability. Because this is a reactive rule it deals with a generic social situation, where the ToM module would reason about the concrete circumstances.

This rule leads PPF to genrate a neutral expression action to suppress the angry expressive behaviour already dispatched. The PPF sends this to the Action module where it returns to the Virtual Body and is dispatched as expressive behaviour. When it returns to Instrinsic Event but its IE tag shows that it has been dealt with, preventing an infinite loop.

5.3 Deliberative Level

The IE component has also passed the emotion to the ToM module which is able to reason about the actual impact of an angry expression on the current goal of the traveller. ToM runs a copy of the appraisal system to assess the impact of the traveller's angry emotion on the bar agents, taking the SI of the traveller and the bar agents into account.

The ToM system will assess the angry expression as reducing the liking of the the bar agents for the traveller, reducing its SI and increasing the threat to its goal of finding the hotel.

This information is passed back to the deliberative system whose planner assesses the ask-the-barman subgoal as a way of achieving the find-hotel goal and deals with the SI threat by proposing a smile and agreement with the proposal of the bar agents. These actions are dispatched to the virtual body but as the smile represents a social signal rather than an affective change it is not passed back round the mind-body-mind loop by the Expressive Behaviour Component but only dispatched for execution.

The net effect on the traveller agent's behaviour is thus an angry micro-expression, suppressed quickly by a neutral expression and then replaced by a social smile.

6 Conclusions

In this paper we have tried to reformulate the architecture of an embodied social agent to support the capability of expressive behaviour as a social signal rather than only as an indicator of the agent's internal state. The first question one should ask is of course whether this is worth doing. There are after all valid arguments for using an embodied agents expressive behaviour to reveal internal state, especially in the case of robots, which are heavy pieces of metallic machinery that should share a human social space in a way that is comfortable for humans. Knowing how a robot is responding and what it is about to do are useful aspects of human-robot interaction. Indeed there is work [76] suggesting that *action-expression* revealing the motivation and context for an agent action through expressive behaviour is a necessity for smooth interaction.

As with so many questions in research, whether this is worth doing is a case of *it depends*. Primarily it depends on the social context in which the embodied agent is expected to perform. We have already seen above that there are applications for which modified expressive behaviour is not only interesting but essential. This seems particularly true for those in which an embodied social agent is following a vocational role, such as tutor, trainer, guide, receptionist. It would be even more true if the application domain related to drama and not to more naturalistic interaction, not only in entertainment applications but also in areas such as role-play based education and training. Here the activity itself is limited by the complexity and expense of supplying human actors, and there is clear scope for the use of social agents.

The approach discussed also supports in a principled way expressive behaviour which is difficult to generate - as against hardcode - without it. One group of such behaviours involves a combination of physiological signals with more cognitively-generated behaviour. Embarrassment, signalled by blushing (a physiological reaction) plus glancing away, would be an example of this. The blush can be generated very rapidly by the intrinsic event raised by the simulated body, while the glance-away is generated later by consciously 'feeling' the emotion as it progresses further through the mind-body-mind loop. A second group of behaviours relate to the overlay of one expression by another as a socially determined expression fails to completely override an internally generated emotion. This supports the known issues with smiles, which often combine with elements of other facial expressions, such as the *disgust* hardcoded in by [40]. This is achieved by a slow decay on a high-intensity emotion dispatched from the simulated body and an overlaid smile from the cognitive stage of the mind-body-mind loop.

Much of the discussion above - very much in line with the literature - has taken a naturalistic approach, using normal human social behaviour as a yardstick. However this assumption should on occasion be challenged. It is not a foregone conclusion that this is the way to incorporate an embodied social agent into everyday human environments. These are agents, they do not and will not for the foreseeable future have human-level abilities given the extreme difficulties involved. It could be that drama rather than naturalism is the more useful paradigm. Indeed the idea of action-expression [76] is more closely related to drama than to naturalism. By showing a sequence of expressions as expressive behaviour is modified, one supplies the human interaction partner with information about the social adjustment of the agent, much in the style of drama, where double-takes and slow realisations are very much standard tropes.

A further argument in favour of a machinery for modifying expressive behaviour is its use in decoding the expressive behaviour of human interaction partners. The problems associated with facial expression recognition have not been the subject of this discussion, but one of the most significant is moving from sensor-based detection of facial movements to an identification of the social signal being deployed. An agent that has no concept of facial expressions as social signals, and works on the basis that there will be a one-to-one mapping between the expression and the users affective state is unlikely to be successful. As argued at the start of this paper, one can recognise a smile, but the signal the smile represents is a different matter. An agent that has a simulation ToM implementation can at least run its own architecture as a decoding mechanism in the current social context.

6.1 Limitations

The most obvious limitation in the discussion of this paper is that it is conceptual and has not yet been implemented. However, though much of the necessary basis for an implementation already exists in the FAtiMA architecture, the main point being made here is that research into embodied social agents should move from a widespread view of agent expressive behaviour as transparently affective especially in the case of facial expressions - and move to the social signal paradigm. We would argue that this also means a move away from the individualistic assumptions underlying many agent architectures into a more socially located account.

A limitation of the suggested changes above in the FAtiMA architecture is that this version of the architecture is entirely symbolic in representation, making truly dynamic expressive behaviour problematic. In order to implement the dynamic PAD-space control of expressive behaviour

discussed in Section 3 one would have to choose the FAtiMA variant FAtiMA-PSI [[77] system discussed above in Section 2.2. Here the FAtiMA symbolically-encoded reactive system is replaced by the PSI [37] five drives: Energy, Integrity, Affiliation, Certainty and Competence and a homeostatic mechanism that chooses actions and goals according to which drives need to be returned to their comfort-zone.

As an example of a non-symbolic approach that has no explicit representation of affect, it is easy to see how it can drive expressive behaviour dynamically. It is less easy to see how one could incorporate the reactive elements of the Social Importance Model. While we argue that most of the discussion is relatively independent of the actual implementation architecture, it is clear that this difficulty would apply to other non-symbolic architectures, such as those generated by machine learning approaches.

The recent ALEXA challenge [78], involving a disembodied (voice-only) conversational agent indicates the most likely solution. The systems that did best in an unrestricted conversational context were compound ones in which machine-learning derived transition networks sat underneath symbolic rule-based systems that provided context and a degree of sanity check. It seems plausible that a compound of this type could supply fast dynamic facial expressions from its sub-symbolic processing, use PAD space to translate these into symbolic representations that are then passed into symbolic SI rules, and pass the outcome back through PAD space into the non-symbolic system.

In conclusion, these are the generic requirements for expressive behaviour as social signals outlined here.

1) A mind-body-mind loop that allows an agent to feel its affect and can trigger.
2) Intrinsic events differentiated from extrinsic events coming from the surrounding environment.
3) A re-evaluation process that responds to intrinsic events and modifies expressive behaviour
4) A model of social interaction that can be used by the re-evaluation system to translate from desired social signal to modified expressive behaviour
5) A ToM that can assess the social impact of an agents expressive behaviour and support a deliberative processing level in modification.

We hope that this paper will help to stimulate work in improving expressive behaviour and changing the default approach to one of social signal generation.

Acknowledgments

The authors would like to thank members of the project ECUTE (ICT-5-4.2257666), whose work was partially supported by the European Commission (EC). The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, which is not responsible for any use that might be made of data appearing therein.

References

[1] P. Ekman and W. V. Friesen, "The Facial Action Coding System," *Consulting*, 1978.
[2] A. Mehrabian, *Nonverbal Communication*. Evanston, IL, USA: Routledge, 2017.
[3] M. F. Jung, "Affective Grounding in Human-Robot Interaction," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction*, 2017, pp. 263–273.
[4] P. M. Cole, "Children's spontaneous control of facial expression," *Child Develop.*, vol. 57, pp. 1309–1321, 1986.
[5] C. Breazeal, "Role of expressive behaviour for robots that learn from people," *Philosophical Trans. Roy. Soc. B: Biological Sci.*, vol. 364, pp. 3527–3538, 2009.
[6] T. Bickmore, "Relational agents: Effecting change through human-computer relationships," *SciencesNew York*, 2003.
[7] L. Hall, C. Hume, S. Tazzyman, A. Deshmukh, S. Janarthanam, H. Hastie, R. Aylett, G. Castellano, F. Papadopoulos, A. Jones, L. J. Corrigan, A. Paiva, P. A. Oliveira, T. Ribeiro, W. Barendregt, S. Serholt, and A. Kappas, "Map reading with an empathic robot tutor," in *Proc. 11th ACM/IEEE Int Conf. Human-Robot Interaction*, 2016, pp. 567–567.
[8] R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. C. Schultz, and J. Wang, "Designing robots for long-term social interaction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2005, pp. 1338–1343.
[9] P. Ekman and W. V. Friesen, "Unmasking the face," *Ann. Phys.*, 1975.
[10] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral Brain Sci.*, 1978.
[11] C. M. D. Melo, P. Carnevale, and J. Gratch, "The Effect of Expression of Anger and Happiness in Computer Agents on Negotiations with Humans," in *Proc. 10th Int. Conf. Auton. Agents Multiagent Syst.*, 2011, pp. 937–944.
[12] R. Hoegen, G. Stratou, and J. Gratch, "Incorporating emotion perception into opponent modeling for social dilemmas," in *Proc. 16th Conf. Auton. Agents MultiAgent Syst.*, 2017, pp. 801–809.
[13] Z. Ambadar, J. F. Cohn, and L. I. Reed, "All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous," *J. Nonverbal Behavior*, vol. 33, pp. 17–34, 2009.
[14] P. M. Niedenthal, M. Mermillod, M. Maringer, and U. Hess, "The Simulation of Smiles (SIMS) model: Embodied simulation and the meaning of facial expression," *Behavioral Brain Sci.*, vol. 33, pp. 417–433, 2010.
[15] P. Ekman and W. V. Friesen, *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. Ishk, 2003.
[16] C. Ritter and R. Aylett, "The partial poker-face," in *Proc. Int. Conf. Intell. Virtual Agents*, 2015, pp. 479–482.
[17] J. Gratch, Z. Nazari, and E. Johnson, "The Misrepresentation Game: How to win at negotiation while seeming like a nice guy," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, 2016, pp. 728–737.
[18] B. M. DePaulo, S. E. Kirkendol, D. A. Kashy, M. M. Wyer, and J. A. Epstein, "Lying in everyday life," *J. Personality Soc. Psychology*, vol. 70, pp. 979–995, 1996.
[19] A. J. Fridlund, *Human Facial Expression*. New York, NY, USA: Academic, 1994.
[20] P. Ekman, "Should we call it expression or communication?" *Innovation: Eur. J. Soc. Sci. Res.*, 1997.
[21] G. Horstmann, "What do facial expressions convey: Feeling states, behavioral intentions, or action requests?" *Emotion*, vol. 3, pp. 150–166, 2003.
[22] A. Scarantino, "Do emotions cause actions, and if so how?" *Emotion Rev.*, 2017.
[23] L. R. Goldberg, "An Alternative "Description of personality": The big-five factor structure," *J. Personality Soc. Psychology*, vol. 59, pp. 1216–1229, 1990.
[24] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotiions*. Cambridge, U.K.: Cambridge Univ. Press, 1988.
[25] A. S. Manstead and A. H. Fischer, "Social appraisal: The social world as object of and influence on appraisal processes," in *Appraisal Processes in Emotion: Theory, Methods, Research*. London, U.K.: Oxford Univ. Press, 2001.
[26] B. Parkinson, A. H. Fischer, and A. S. Manstead, *Emotion in Social Relations: Cultural, Group, and Interpersonal Processes*. Evanston, IL, USA: Routledge, 2004.
[27] R. S. Lazarus and S. Folkman, "Transactional theory and research on emotions and coping," *Eur. J. Personality*, 1987.
[28] J. Dias and A. Paiva, "Feeling and reasoning: A computational model for emotional characters," in *Proc. Portuguese Conf. Artif. Intell.*, 2005, pp. 127–140.
[29] J. Dias, R. Aylett, A. Paiva, and H. Reis, "The great deceivers: Virtual agents and believable lies," in *Proc. Annu. Meeting Cognitive Sci. Soc.*, 2013, vol. 35, no. 35.

[30] S. Mascarenhas, J. Dias, R. Prada, and A. Paiva, "A dimensional model for cultural behavior in virtual agents," *Appl. Artif. Intell.*, vol. 24, no. 6, pp. 552–574, 2010.

[31] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive Developmental Robotics: A Survey," *IEEE Trans. Auton. Mental Develop.*, vol. 1, no. 1, pp. 12–34, May 2009.

[32] D. Vernon, "Enaction as a conceptual framework for developmental cognitive robotics," *Paladyn, J. Behavioral Robot.*, vol. 1, pp. 89–98, 2010.

[33] A. Watanabe, M. Ogino, and M. Asada, "Mapping facial expression to internal states based on intuitive parenting," *J. Robot. Mechatronics*, vol. 19, 2007.

[34] K. R. Scherer, "Appraisal considered as a process of multilevel sequential checking," 2001.

[35] M. Courgeon, C. Clavel, and J.-C. Martin, "Appraising emotional events during a real-time interactive game," in *Proc. Int. Workshop Affective-Aware Virtual Agents Soc. Robots*, 2009, Art. no. 7.

[36] L. Canamero, "Issues in the design of emotional agents," *Cybern. Syst.*, 2001.

[37] D. Dörner and C. D. Güss, "PSI: A computational architecture of cognition, motivation, and emotion," *Rev. General Psychology*, vol. 17, no. 3, 2013, Art. no. 297.

[38] M. Y. Lim and R. Aylett, "An emergent emotion model for an affective mobile guide with attitude," *Appl. Artif. Intell.*, vol. 23, no. 9, pp. 835–854, 2009.

[39] S. Marsella and J. Gratch, *EMA: A Computational Model of Appraisal Dynamics*. 2006.

[40] M. Ochs, R. Niewiadomski, C. Pelachaud, and D. Sadek, "Intelligent expressions of emotions," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2005, pp. 707–714.

[41] R. Niewiadomski, M. Ochs, and C. Pelachaud, "Expressions of empathy in ECAs," in *Proc. 8th Int. Conf. Intell. Virtual Agents*, 2008, pp. 37–44.

[42] S. W. McQuiggan and J. C. Lester, "Modeling and evaluating empathy in embodied companion agents," *Int. J. Human Comput. Stud.*, vol. 65, pp. 348–360, 2007.

[43] J. Robison, S. McQuiggan, and J. Lester, "Evaluating the consequences of affective feedback in intelligent tutoring systems," in *Proc. 3rd Int. Conf. Affective Comput. Intell. Interaction Workshops*, 2009, pp. 1–6.

[44] S. C. Marsella, W. L. Johnson, and C. M. Labore, "Interactive Pedagogical Drama for Health Interventions," in *Proc. 11th Int. Conf. Artif. Intell. Educ.*, 2003.

[45] C. M. De Melo, P. Carnevale, and J. Gratch, "The effect of virtual agents' emotion displays and appraisals on people's decision making in negotiation," in *Proc. Int. Conf. Intell. Virtual Agents*, 2012, pp. 53–66.

[46] C. M. D. Melo, P. Carnevale, S. J. Read, and J. Gratch, "Bayesian model of the social effects of emotion in decision-making in multiagent systems," in *Proc. 11th Int. Conf. Auton. Agents Multiagent Syst.*, 2012, pp. 55–62.

[47] M. Rehm and E. André, "Catch me if you can exploring lying agents in social settings," *Aamas*, 2005.

[48] P. Ekman, "Darwin, Deception, and Facial Expression," *Ann. New York Acad. Sci.*, vol. 1000, pp. 205–221, 2003.

[49] B. De Carolis, C. Pelachaud, I. Poggi, and F. De Rosis, "Behavior planning for a reflexive agent," in *Proc. 17th Int. Joint Conf. Artif. Intell.- Vol. 2*, 2001, pp. 1059–1064.

[50] A. Rao and M. Georgeff, "Modeling rational agents within a BDI-architecture," in *Readings in Agents*. San Francisco, CA, USA: Morgan Kaufmann, 1997.

[51] D. Pereira, E. Oliveira, N. Moreira, and L. Sarmento, "Towards an architecture for emotional BDI agents," in *Proc. Portuguese Conf. Artif. Intell.*, 2005, pp. 40–46.

[52] I. Poggi and C. Pelachaud, "Emotional meaning and expression in animated faces," in *Proc. Int. Workshop Affective Interactions*, 2000, pp. 182–195.

[53] C. Castelfranchi and Y. H. Tan, "The role of trust and deception in virtual societies," *Int. J. Electron. Commerce*, 2002.

[54] B. DeCarolis, C. Pelachaud, I. Poggi, and M. Steedman, "APML, a mark-up language for believable behavior generation," *Lifelike Characters Tools Affective Functions Appl.*, 2004.

[55] M. Schröder, "The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems," *Adv. Human-Comput. Interaction*, vol. 2010, 2010, Art. no. 2.

[56] C. Becker-Asano and H. Ishiguro, "Evaluating facial displays of emotion for the android robot Geminoid F," in *Proc. Workshop Affective Comput. Intell.*, 2011, pp. 1–8.

[57] A. R. Damasio, *Descartes' Error*. New York, NY, USA: Random House, 2006.

[58] K. R. Scherer, "On the nature and function of emotion: A component process approach," *Approaches Emotion*, vol. 2293, 1984, Art. no. 317.

[59] P. Gebhard, T. Schneeberger, T. Baur, and E. André, "MARSSI: Model of appraisal, regulation, and social signal interpretation," in *Proc. 17th Int. Conf. Auton. Agents MultiAgent Syst.*, 2018, pp. 497–506.

[60] O. Lemon and O. Pietquin, "Machine Learning for Spoken Dialogue Systems," in *Proc. INTERSPEECH*, 2007.

[61] S. G. Shamay-Tsoory, J. Aharon-Peretz, and D. Perry, "Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions," *Brain*, vol. 132, pp. 617–627, 2009.

[62] O. Lemon, A. Bracy, A. Gruenstein, and S. Peters, "The WITAS multi-modal dialogue system I," in *Proc. 7th Eur. Conf. Speech Commun. Technol.*, 2001.

[63] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson, "Towards a common framework for multimodal generation: The behavior markup language," in *Proc. Int. Workshop Intell. Virtual Agents*, 2006, pp. 205–217.

[64] M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann, "SmartBody: Behavior realization for embodied conversational agents," in *Proc. 7th Int. Joint Conf. Auton. Agents Multiagent Syst*, 2008, pp. 151–158.

[65] T. Ribeiro, M. Vala, and A. Paiva, "Thalamus: Closing the mind-body loop in interactive embodied characters," in *Proc. Int. Conf. Intell. Virtual Agents*, 2012, pp. 189–195.

[66] A. Mehrabian, "Analysis of the big-five personality factors in terms of the PAD temperament model," *Australian J. Psychology*, vol. 48, no. 2, pp. 86–92, 1996.

[67] E. B. Roesch, L. Tamarit, L. Reveret, D. Grandjean, D. Sander, and K. R. Scherer, "FACSGen: A tool to synthesize emotional facial expressions through systematic manipulation of facial action units," *J. Nonverbal Behavior*, vol. 35, pp. 1–16, 2011.

[68] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personality Soc. Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[69] R. E. Jack, O. G. B. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proc. Nat. Acad. Sci.*, vol. 109, pp. 7241–7244, 2012.

[70] L. Malatesta, A. Raouzaiou, K. Karpouzis, and S. Kollias, "Towards modeling embodied conversational agent character profiles using appraisal theory predictions in expression synthesis," *Appl. Intell.*, vol. 30, pp. 58–64, 2009.

[71] P. E. McKenna, M. Y. Lim, A. Ghosh, R. Aylett, F. Broz, and G. Rajendran, "Do you think I approve of that? Designing facial expressions for a robot," in *Proc. Int. Conf. Soc. Robot.*, 2017, pp. 188–197.

[72] S. Mascarenhas, N. Degens, A. Paiva, R. Prada, G. J. Hofstede, A. Beulens, and R. Aylett, "Modeling culture in intelligent virtual agents," *Auton. Agents Multi-Agent Syst.*, vol. 30, no. 5, pp. 931–962, 2016.

[73] T. D. Kemper, *Status, Power and Ritual Interaction: A Relational Reading of Durkheim, Goffman and Collins*. Evanston, IL, USA: Routledge, 2016.

[74] R. Aylett, L. Hall, S. Tazzyman, B. Endrass, E. André, C. Ritter, A. Nazir, A. Paiva, G. Höfstede, and A. Kappas, "Werewolves, cheats, and cultural sensitivity," in *Proc. Int. Conf. Auton. Agents Multi-Agent Syst.*, 2014, pp. 1085–1092.

[75] C. Ritter and R. Aylett, "The Partial Poker-Face," in *Proc. Int. Conf. Intell. Virtual Agents*, 2015, pp. 479–482.

[76] P. Sengers, "Do the thing right: An architecture for action-expression," in *Proc. 2nd Int. Conf. Auton. Agents*, 1998, pp. 24–31.

[77] M. Y. Lim, J. Dias, R. Aylett, and A. Paiva, "Creating adaptive affective autonomous NPCs," *Auton. Agents Multi-Agent Syst.*, vol. 24, no. 2, pp. 287–311, 2012.

[78] A. Ram, R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, and others, "Conversational AI: The science behind the Alexa prize," *arXiv:1801.03604*, 2018.

Ruth Aylett became a professor of Computer Science at Heriot-Watt University in 2004 where she is a member of the Edinburgh Centre for Robotics and researches social agents, human-robot interaction and affective computing. She currently leads a UKRC-funded project SoCoRo investigated the development of a robot trainer in social signal recognition for high-functioning adults with an ASD.

Christopher Ritter received the German diploma in computer sciences from Friedrich-Alexander University Erlangen-Nuremberg. He is currently working toward the doctoral degree in computer sciences at the University of Bielefeld, Social Cognitive Systems group, CITEC.

Mei Yii Lim received the PhD degree from Heriot-Watt University, in 2007. She has worked as a post-doc researcher on EU-funded projects eCIRCUS, eCUTE, SOCIETIES, LIREC and EMOTE. She is currently a researcher on the UK-funded project SoCoRo.

Frank Broz received the PhD degree in robotics from the Carnegie-Mellon University Robotics Institute. He is an assistant professor of computer science with Heriot-Watt University. He was a senior research fellow with Plymouth University working on the Robot-ERA project before joining Heriot-Watt University in 2015 and researches artificial intelligence, human-robot interaction and social robotics.

Peter E McKenna received the doctorate degree in psychology from Heriot-Watt University, Edinburgh. He is currently a research associate on the EPSRC funded SoCoRo project - he and the team are developing a socially competent robot to teach adults with an ASD social and employment skills.

Ingo Keller received the diploma in computer science from Technische Universitt Dresden (TUD), Germany, in 2010. He joined Heriot-Watt University in 2014 to pursue the PhD degree and is researching interactive object learning in the area of Teachable Robots. He is also investigating aspects of gesture synthesis in the SoCoRo project.

Gnanathusharan Rajendran received the PhD degree in developmental psychology from the University of Nottingham. He is an associate professor in psychology with Heriot-Watt University. He joined Heriot-Watt University in 2012 as a reader and specialises in typical and atypical cognitive and social development, digital education, and social robotics.